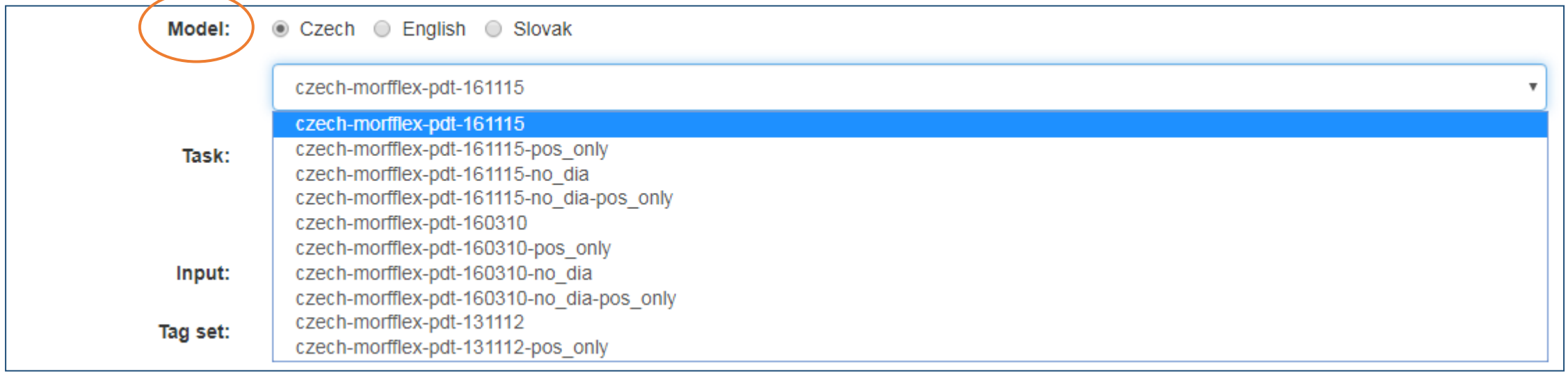
Lehký úvod do strojového učení

Barbora Hladká
ÚFAL MFF UK

Motivace

- Spuštění lingvistických nástrojů https://lindat.cz/services vyžaduje mj. i výběr modelu. Co je to model?



- Proč nástroje chybují?

Motivace
Naučme počítač tvaroslovný rozbor

1. Uvažujeme slovní druhy a mluvnické kategorie a jejich hodnoty.

Slovní druhy

- podstatná jména (N)
- přídavná jména (A)
- zájmena (P)
- číslovky (C)
- slovesa (V)
- příslovce (D)
- předložky (R)
- spojky (J)
- částice (T)
- citoslovce (I)

Mluvnické kategorie

- rod (M,I, …)
- číslo (S,P,…)
- pád (1-7)
- osoba (1-3)
- čas (P,R,…)
- …

Naučme počítač tvaroslovný rozbor

2. Zavedeme značky

slovní druh | rod | číslo | pád | osoba | čas | …

3. Vytvoříme učebnici

Satoranský	NMS1--…
řádil	VMS-3P…
v	R--6-- …
přípravě	NFS6--…
v	R--6-- …
Torontu	NNS6--…
.	Z-----…
…	…

Naučme počítač tvaroslovný rozbor

3. Převedeme učebnici do frekvenčních tabulek

slovo	se značkou	kolikrát
Satoranský	NMS1--	10x
měl	VMS-3P	15x
přípravě	NFS6--	10x
přípravě	NFS3--	8x
příležitost	NFS4--	10x
příležitost	NFS1--	5x
se	R--7--	40x
dává	VMS-3R	20x
dává	VFS-3R	20x

dvojice značek	za sebou	kolikrát
NMS1---	VMS-3R	10x
VMS-3R	R--7–	15x
R--7–	NMS7--	2x

Naučme počítač tvaroslovný rozbor

Satoranský	NMS1--	10x
měl	VMS-3P	15x
přípravě	NFS6--	10x
přípravě	NFS3--	8x
…		
příležitost	NFS4--	10x
příležitost	NFS1--	5x
…		
se	R--7--	40x
dává	VMS-3R	1x
dává	VFS-3R	1x
NMS1---	VMS-3R	10x
VMS-3R	R--7–	15x
R--7–	NMS7--	2x

Satoranský NMS1--

dává VMS-3R

koš R--7--

koš a předložka?

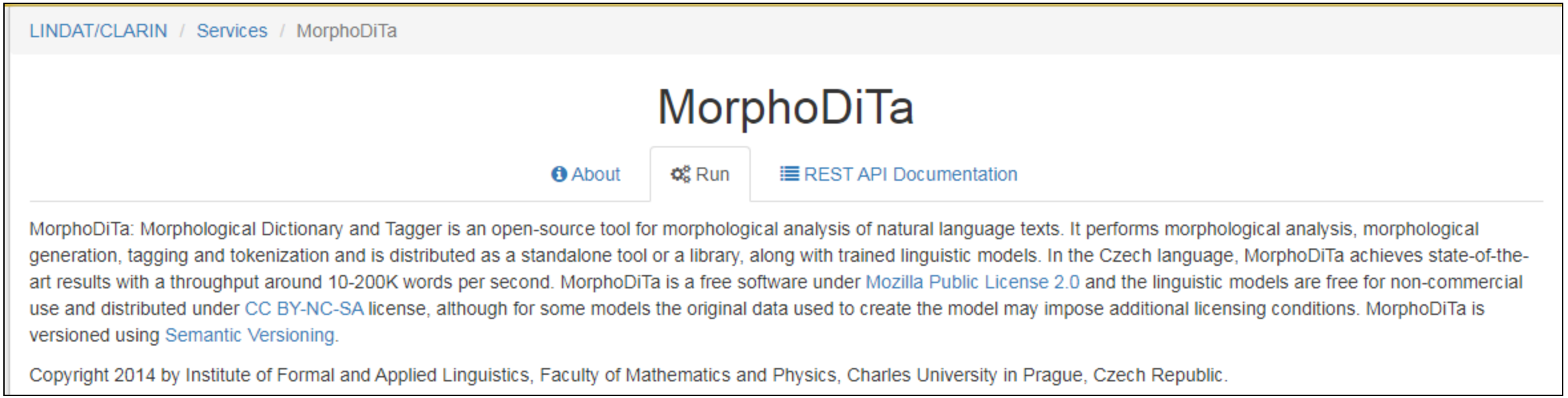
4. Satoranský dává koš.

Toto je strojové učení! S učitelem.

Vytvoříme příklady a navedeme počítač, jak se z nich učit.

MorphoDiTa umí tvaroslovný rozbor

https://lindat.mff.cuni.cz/en/services#MorphoDiTa



Klíčové pojmy

- zlatá data jsou příklady (v motivačním příkladu učebnice)
    - rozdělení na trénovací a testovací množiny
- reprezentace dat je popis příkladů pomocí atributů

  (v mot. příkladu dvojice slovo, značka a značka, značka)
- model je způsob reprezentace dat + trénovací množina (v mot. příkladu frekvenční tabulky)
- metoda učení je způsob učení z modelu

  (v mot. příkladu způsob procházení frekvenční ch tabulek)
- trénování modelu je metoda učení + model
- evaluace je vyhodnocení úspěšnosti modelu na testovací množině

MorphoDiTa
Zlatá data

Prague Dependency Treebank 3.5

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

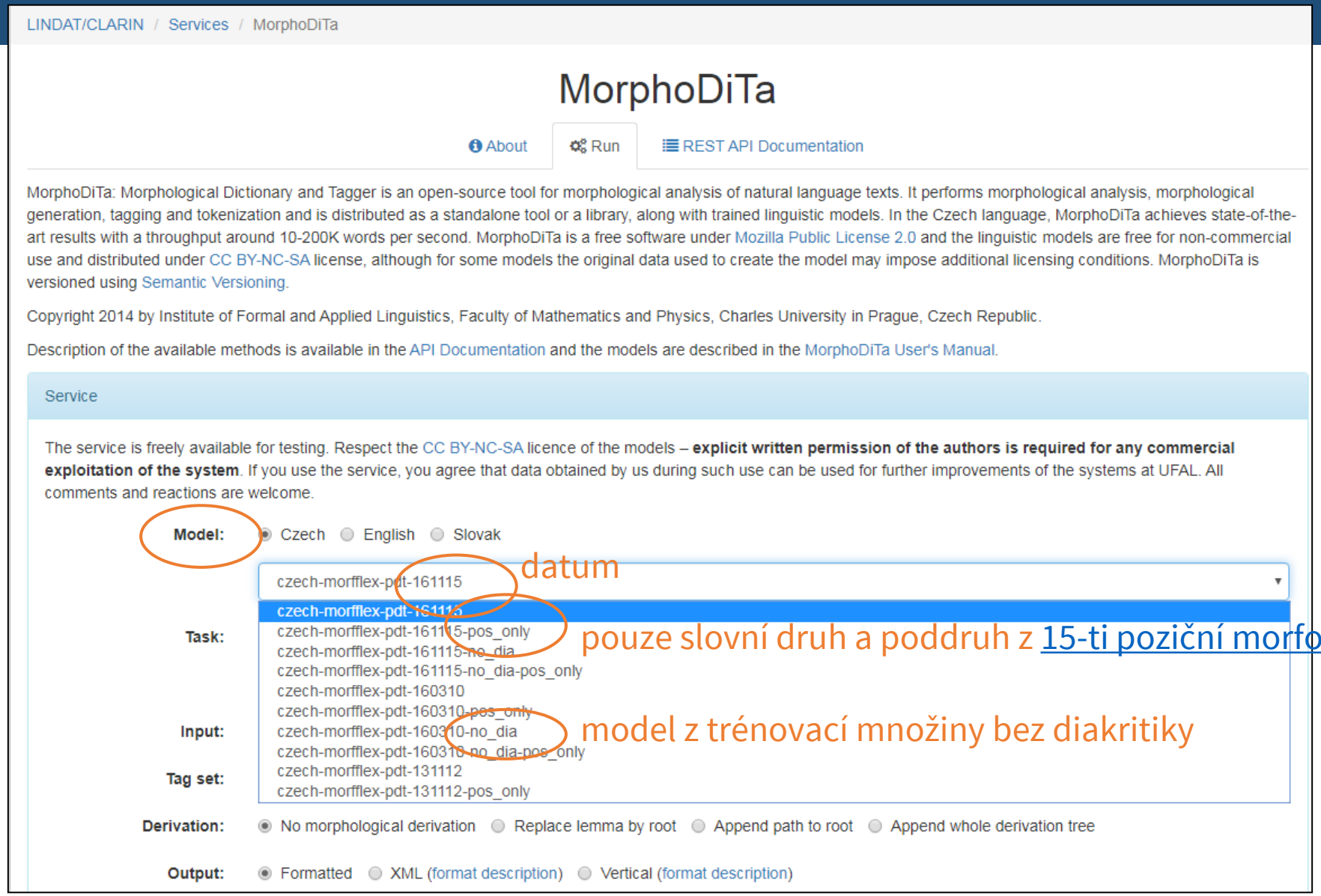
Hajič, Jan; Bejček, Eduard; Bémová, Alevtina; et al., 2018, *Prague Dependency Treebank 3.5*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2621.

Share:

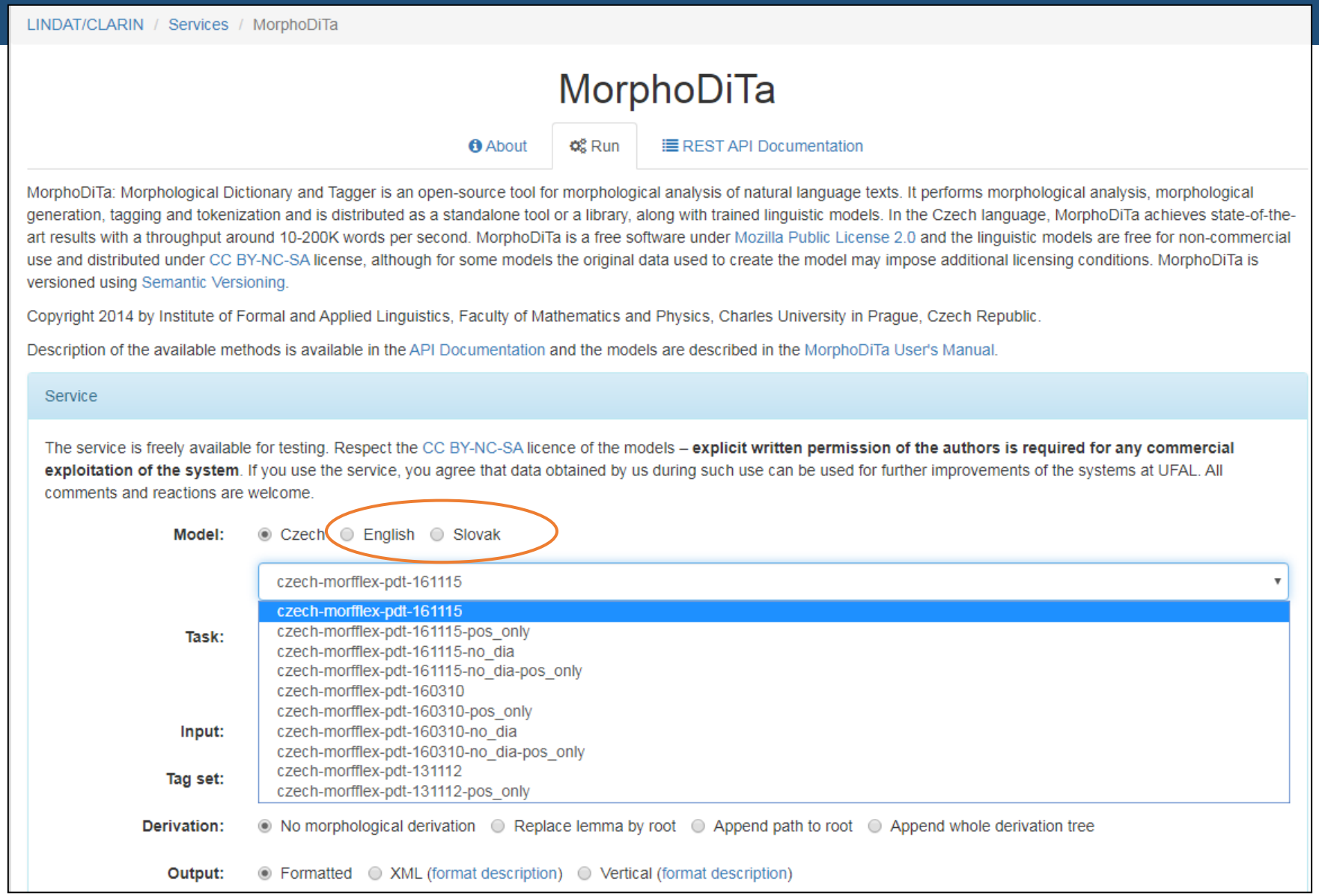
LINDAT / CLARIN

Authors	Hajič, Jan ; Bejček, Eduard ; Bémová, Alevtina ; Buráňová, Eva ; Hajičová, Eva ; Havelka, Jiří ; Homola, Petr ; Kárník, Jiří ; Kettnerová, Václava ; Klyueva, Natalia ; Kolářová, Veronika ; Kučová, Lucie ; Lopatková, Markéta ; Mikulová, Marie ; Mírovský, Jiří ; Nedoluzhko, Anna ; Pajas, Petr ; Panevová, Jarmila ; Poláková, Lucie ; Rysová, Magdaléna ; Sgall, Petr ; Spoustová, Johanka ; Straňák, Pavel ; Synková, Pavlína ; Ševčíková, Magda ; Štěpánek, Jan ; Urešová, Zdeňka ; Vidová Hladká, Barbora ; Zeman, Daniel ; Zikánová, Šárka ; Žabokrtský, Zdeněk
Item identifier	http://hdl.handle.net/11234/1-2621
Project URL	http://ufal.mff.cuni.cz/pdt3.5
Demo URL	http://ufal.mff.cuni.cz/pdt3.5
Date issued	2018-02-19
Type	corpus
Size	49431 sentences, 2000000 words
Language(s)	Czech
Description	The Prague Dependency Treebank 3.5 is the 2018 edition of the core Prague Dependency Treebank (PDT). It contains all PDT annotation made at the Institute of Formal and Applied Linguistics under various projects between 1996 and 2018 on the original texts, i.e., all annotation from PDT 1.0, PDT 2.0, PDT 2.5, PDT 3.0, PDiT 1.0 and PDiT 2.0, plus corrections, new structure of basic documentation and new list of authors covering all previous editions. The Prague Dependency Treebank 3.5 (PDT 3.5) contains the same texts as the previous versions since 2.0; there are 49,431 annotated sentences (over 800 thousand nodes) on all layers, from tectogrammatical to words, and additional sentences on the analytical (surface dependency syntax) and morphological layers of annotation (approx. 2 million words in total). Closely linked to the tectogarammtical layer is the annotation of sentence information structure, multiword expressions, coreference, bridging relations and discourse relations.

MorphoDiTa Modely



MorphoDiTa
Modely



metody učení jsou jazykově nezávislé, jazykově závislá jsou zlatá data

Statistické strojové učení

- Zlatá data jsou výběrem z populace příkladů. Proto můžeme trpět
    - nedostatkem příkladů
    - v testovací množině jsou příklady, které nejsou v trénovací

      (v motivačním příkladu koš)
    - v trénovací množině jsou příklady s nízkou četností
- Z toho pramení chyby modelů

LINDAT CLARIAH-CZ

podzimní tutoriály

Národní galerie Praha - Veletržní palác
Dukelských hrdinů 47, Praha 7
Auditorium, 6. patro

tutoriál
17.10.2019
9-13 hod.

Praktická část

09:00-11:00 Nástroje pro zpracování textů
- Webové služby https://lindat.cz/services
- Lehký úvod do strojového učení
- Demonstrace spuštění vybraných nástrojů

Prezentační část

11:30-12:00 Bibliografie dějin českých zemí
12:00-13:00 Diskuse

- Otevřeno všem zájemcům
- Registrace není nutná a účast je bezplatná
- Každý účastník tutoriálu obdrží dárek

Výzkumná infrastruktura LINDAT/CLARIAH-CZ

(LM2018101 a CZ.02.1.01/0.0/0.0/16_013/0001781; původně LM2010013, LM2015071)

je financována Ministerstvem školství, mládeže a tělovýchovy České republiky

v programu LM „Velké infrastruktury".